Dimensions of L2 Oral Language Performance: A Study of Complexity, Accuracy, and Fluency Development Over Time

Miho Tomita,
Graduate School of Languages and Linguistics, Sophia University, Japan

This study examined the developmental patterns of second language (L2) oral language performance as measured by complexity, accuracy, and fluency (CAF) and the relationship between these three variables over time. A total of 31 Japanese-L1 university students, grouped into two proficiency levels (lower and intermediate), participated in a semester-long task-based speaking course. Speaking tests involving impromptu speech tasks were administered four times over the 15-week semester, and learners' oral data were analyzed to measure CAF. The results indicated that syntactic complexity had mild growth over time, with some fluctuations. Lexical complexity showed a mild U-shaped curve with slight changes in growth. Accuracy showed U-shaped trajectories, showing a decline followed by a steeper increase over time, and fluency exhibited steady growth over time. Regarding correlations between CAF, trade-off effects were evident between lexical complexity and syntactic complexity and between lexical complexity and accuracy. We observed a positive correlation between accuracy and syntactic complexity and between fluency and syntactic complexity. Regarding the relation between fluency and accuracy, the results were mixed, and there was an observed trend towards significance between fluency and lexical complexity. The findings also indicated that lower- and intermediate-proficiency learners had similar change trajectories except for one syntactic complexity measure.

Keywords: CAF, Complexity, Accuracy, Fluency, Speaking

1. Introduction

Applied linguists use three components to evaluate second language (L2) development and proficiency; complexity, accuracy, and fluency (CAF) (e.g., Ellis, 2003; Ellis & Barkhuizen, 2005; Skehan, 1998); which, when taken together, reveal the learner's L2 mastery. Although the weighting of each component depends on learning goals and other factors, L2 lessons should be planned to allow learners to improve all three components equally. However, the developmental patterns of the CAF components are complex and learners cannot devote equal attention to every aspect (e.g., Bamanger & Gashan, 2014; Foster & Skehan, 1996; Robinson, 2003; Sasayama & Izumi, 2012). Stronger performance in one component may correspond to worse performance in another.

Skehan's (1996, 1998) limited capacity hypothesis observes that human attentional capacity is limited, so learners must choose one aspect of CAF to prioritize. This hypothesis predicts competitive relations in CAF that would prevent all three aspects from improving simultaneously. Robinson (2011) challenged this perspective with his cognitive hypothesis, arguing for a multiple-resource attentional model in which learners can access multiple attentional pools that are not in competition. Thus, learners could simultaneously improve complexity and accuracy at the expense of fluency. Skehan's and Robinson's hypotheses have been tested (e.g., Foster & Skehan, 1997; Yuan & Ellis, 2003), and trade-off effects have been identified in CAF.

However, most studies have only investigated language performance at a single time within a homogenous proficiency group. In such studies, it is difficult to observe how the aspects of CAF may change. CAF may change over time as proficiency increases, and the multiple results regarding trade-off effects could be better explained using observations of different relationships at different levels of proficiency. Cross-sectional studies of oral data from learners of differing proficiencies are often used as substitutes for longitudinal studies;

however, the efficacy of this is debatable (Larsen-Freeman & Long, 1991). Thus, studies with repeated measures are necessary to better understand CAF development (Norris & Ortega, 2009).

This study describes the development of oral performance among Japanese L2 learners in CAF. How do the aspects of CAF develop over time and interact as they develop? Do they compete? Do complexity and accuracy develop simultaneously? Is there a trade-off in CAF during L2 development? Can CAF grow without competition? Finally, is the developmental pattern complex, that is, instead of linear growth or a straightforward pattern, does the trajectory diverge from this (e.g., taking a U-shaped or zig-zag path)?

In this study, 31 Japanese-L1 university students were divided into two proficiency levels and participated in a semester-long task-based speaking course. Speaking tests (impromptu speech tasks) were administered four times. Students' speeches were recorded, transcribed, coded, and assessed to measure their CAF; these scores were analyzed to determine changes in performance over time. The interaction of CAF components over time and the effects of learner proficiency on CAF development were investigated.

2. Literature Review

2.1 *Cognitive Limitation*

Many L2 development studies have focused on the interrelationships between the CAF components (e.g., Skehan, 1998; Robinson, 2011; Norris and Ortega, 2009). Many researchers accept that learners have limited resources available for improving their CAF performance. Skehan (1996, 1998) held that because learners cannot concentrate on every aspect of CAF at once, their concentration on one draws attention away from another. Thus, if a learner's output becomes more complex, accuracy and fluency may not improve; thus, increased complexity "might be associated with lower fluency, or raised accuracy with lower complexity" (Skehan, 2015, p. 125).

Robinson's (2011) cognition hypothesis proposed a multiple-resources attentional model, where learners would not need to trade gains in attention to one aspect of production against losses in another. In this framework, complexity and accuracy are correlated, and complex tasks can enhance the development of accuracy and complexity. Robinson (2003) argued that greater functional demands of the task lead language learners to pay closer attention to language. Thus, during complex task performance, "learners attempt to achieve greater syntacticization and grammaticization of their current interlanguage" (p. 77) to meet the increased cognitive demand. Although complexity and accuracy may improve together, Robinson (1995) thought that they may not have positive relationships with fluency.

Larsen-Freeman (2006, 2009) considers CAF to be a dynamic system in the form of a set of variables that interact over time, such that language development is a dynamic and complex process. The dynamic systems theory (DST) approach regards language acquisition and development as possessing growth and decline characteristics that are influenced by many internal and external factors (e.g., de Bot, 2007, 2008; de Bot & Larsen-Freeman, 2011): the aspects of CAF develop dynamically and interactively. A change in any one component might affect the others unpredictably.

2.2 *CAF Interaction*

Empirical examinations of trade-off effects have produced inconsistent findings. Bei (2013) reported a strong correlation between fluency and accuracy but competition between accuracy and complexity. Koizumi (2005) found marginal to fairly weak correlations among syntactic complexity, accuracy, and fluency. Koizumi and In'nami (2014) reported moderate or strong positive correlations of syntactic complexity with accuracy and fluency but a weak relationship between accuracy and fluency. Yuan and Ellis (2003) showed that greater structural complexity and fewer error-free clauses appeared at the expense of fluency. However, Michel, Kuiken, and Vedder (2007) found more oral accuracy and lexical complexity, but grammatical complexity and fluency did not improve.

Most work on this subject has used a single time point; researchers working in a DST perspective have conducted longitudinal studies to assess the elements of CAF (using data largely collected from written texts). Verspoor, Lowie, and Van Dijik (2008) observed that lexical and syntactic complexity had a slightly negative correlation. Spoelman and Verspoor (2010) explored writing samples from a single learner over 3 years; they suggested that, although complexity and accuracy showed growth, the development was nonlinear rather than a complex interactional pattern among the three components. Despite this finding, Yang and Sun (2015) showed that the components of CAF, especially lexical complexity and grammatical complexity, were correlated with each other over 10 months.

Ferrari (2012) investigated the oral development of CAF in four L2 learners over 3 years using monologic and dialogic tasks. Although learners' CAF developed, trade-off effects were also observed. Additionally, each learner had a different trajectory and speed of development. By contrast, Vercellotti (2017), in a longitudinal study, found no trade-off effects, finding linear change trajectories for CAF (except for lexical aspects, which were nonlinear) over 6 months, and positive within-individual correlation results. Polat and Kim (2014) studied one uninstructed L2 learner over 12 months and found that lexical complexity increased steadily over time, syntactic complexity increased somewhat, and accuracy did not increase.

Competition among the elements of CAF has been intensively investigated, but most works hitherto have examined performance at a single point in time, not development over time; thus, additional longitudinal studies are necessary. Some researchers (e.g., Ferrari, 2012; Spoelman and Verspoor, 2010) have conducted longitudinal studies of learners' written text (e.g., Alavi & Sadeghi, 2017; Yang & Sun, 2015). Further research into the effects of modality is required to assess differences between the results from the written text and those of oral data. Kuiken and Vedder (2012) compared oral and written data and observed minor differences between the two, but Ellis and Yuan (2005) found differences were observed in all three components in a similar study: in the written data, complexity and accuracy were higher, and fluency was lower. To understand the overall development of learners' production, the progress of all components in relation to proficiency should be assessed. Observations of learners at different proficiency levels may yield differing change trajectories that complement earlier findings. Close observation of these trajectories could enable the assessment of patterns of oral development by proficiency level. Such information could enable decision-making when matching learners at different levels to tasks that suit their development.

The following research questions guided this study:
RQ1: How do the CAF of L2 speaking develop over time?
RQ2: How do the components of CAF interact with each other in the development of L2 speaking over time?
RQ3: How does learners' proficiency relate to CAF improvement in the spoken production of L2 over time?

3. Research Design

3.1 *Participants*

The participants were 31 Japanese EFL students at a private Japanese college[i]. They were in their second year (aged 19–21 years) and were streamed according to their TOEIC L&R scores[ii]. Group 1 was at a lower level, with a mean TOEIC score of 455.4. Group 2 was at an intermediate level, with a mean TOEIC of 626.8. The students in each group met for 90 minutes per week, with 15 meetings per term. Materials were provided by the school. The textbook used featured many preparation questions and exercises for the TOEIC speaking test, which all students had to take at the end of the second year.

3.2 *Speaking Tests*

Data were collected four times in the semester, roughly one month apart. An impromptu speech task was used, developed after an actual TOEIC speaking test. Students' speech was recorded during regular speaking class time in a language media lab. The participants were instructed to speak on a given topic and were given fifteen seconds to plan, following an actual TOEIC speaking test (Table 1). After recording, all data were transcribed by the Author.

Table 1. Speech topics for speaking tests

Round 1	April-17	Some people prefer to live in a small town. Others prefer to live in a big city. Which place would you prefer to live in? Give reasons or examples to support your opinion
Round 2	May-17	Some people prefer to eat at restaurants. Other people prefer to eat at home. Which do you prefer? Give reasons or examples to support your opinion
Round 3	June-17	Do you think cram schools are important? Why? Give reasons or examples to support your opinion
Round 4	July-17	Do you agree or disagree with the following statement. "Students should have part-time jobs." Give reasons or examples to support your opinion

3.3 *Self-reported Evaluation (Questionnaire)*

Beyond the speaking test, a seven-item questionnaire was presented to explore factors that might affect the results, such as individual differences (e.g., fear of making errors and task difficulty perception). Immediately after the test, students completed a self-reported evaluation/questionnaire with responses on a five-point Likert scale, shown in Table 2 (translated by the Author). The questionnaire was developed by the Author, using the Foreign Language Classroom Anxiety Scale (FLCAS) from Horwitz et al. (1986)[iii].

The questionnaire contained seven self-evaluation items on the task, anxiety, and confidence in oral performance. At the bottom of the form was an open-ended comment box where students could provide comments.[iv]

Table 2. Self-evaluation sheet

	strongly disagree	disagree	neither agree or disagree	agree	strongly agree
1. This task was difficult.	1	2	3	4	5
2. I was able to express my opinion.	1	2	3	4	5
3. I paid attention to correct pronunciation while speaking.	1	2	3	4	5
4. I paid attention to correct grammar usage while speaking.	1	2	3	4	5
5. I found myself thinking in Japanese when delivering a speech.	1	2	3	4	5
6. I felt confident when I was speaking English during the task.	1	2	3	4	5
7. I didn't worry about making mistakes during the task.	1	2	3	4	5

3.4 *Analysis*

To answer the research questions, all data were transcribed and coded into clauses (finite and non-finite clauses) and AS-units[v]. Table 3 shows a summary of the measures used to examine the CAF: 1. syntactic complexity 1. (SC1: number of words per AS-unit), 2. syntactic complexity 2 (SC2: number of clauses per AS-unit), 3. lexical complexity (LC: index of lexical diversity (D)), 4. accuracy 1 (A1: number of error-free clause per clause), 5. accuracy 2 (A2: number of errors per word), 6. fluency 1 (F1: number of words per minute), and 7. fluency 2 (F2: number of clauses per minute). During the coding, disfluency markers (e.g., filled pauses such as “uh” or “er,” repetitions, and false starts) were not counted as errors or words.

Table 3. Summary of Seven Measures

Factor	Code	Measure	Source Example
Syntactic Complexity	SC1	No. of words per AS-Unit	Mehnert (1998), Bygate (2001)
	SC2	No. of clauses per AS-Unit	Yuan & Ellis (2003), Koizumi (2005)
Lexical Complexity	LC	Index of lexical diversity (D)	Kormos & Denes (2004), Vecellotti (2017)
Accuracy	AC1	No. of error-free clause per clause	Ellis & Barkhuizen (2005)
	AC2	No. of errors per word	Takiguchi (2004), Koizumi (2005)
Fluency	F1	No. of words per minute	Koizumi & Yamanouchi (2003), Ishikawa (2015)
	F2	No. of clauses per minute	Takiguchi (2004)

Both syntactic and lexical complexity were studied. To measure SC1, the number of words per AS-unit was calculated, following Bygate (2001). The AS-unit was used instead of the T-unit or C-unit because it is more appropriate for measuring the output of lower- or intermediate-level learners (Koizumi, 2005). The number of clauses per AS-unit was chosen to determine SC2 because it relates to the complexity level of syntactic structures (Koizumi, 2005).

D, an index of lexical diversity was used for lexical complexity (Kormos & Denes, 2004). D represents the proportion of content words to total words. This value was chosen because it is assumed to be the most accurate instrument for comparing lexical diversity between texts of different lengths, even relatively short ones (e.g., Malvern et al., 2004; Johansson, 2008; Daller et al., 2003).

As in Ellis and Barkhuizen (2005), to determine the A1 rate, the number of error-free clauses was compared with the total number of clauses without counting the discourse errors

because accuracy was judged as the learner's ability to speak without errors in real-time communication (Wolfe-Quintero et al., 1998). To determine the A2 rate, the number of errors and the error rate per word, considered sensitive accuracy measures, were counted, without counting the discourse errors (Mehnert, 1998). Takiguchi (2004) and Koizumi (2005) used the same measure for their speaking performance accuracy analyses.

As in Takiguchi (2004), Koizumi & Yamanouchi (2003), and Ishikawa (2015), the speed fluency measures, F1, the number of words per minute, and F2, the number of clauses per minute, were examined. Unit length (e.g., clauses per AS-unit) was not employed as a fluency measure, assuming its greater correlation with syntactic complexity than that with fluency (Koizumi, 2005). Further, the number of words per minute is "one of the most reliable and stable measures of L2 speech fluency" (Ishikawa, 2015, p.519). Pause information was not examined in this study as it would have required a specialized tool for fine-grained analysis (Griffiths, 1991).

To observe the CAF trajectories, the mean CAF scores in each round for each group were calculated, and a one-way repeated-measure ANOVA was employed to compare the mean scores and determine the presence of any significant differences between the time points. To answer research question 2, that is, to observe the CAF construct relationships over time, within-individual correlation analyses were conducted. The results from rounds 1 to 4 for each measure were entered into the calculations, and the correlations between the trajectories were analyzed. Within-individual correlation analyses test for the presence of a link between the trajectories within individual development. To examine research question 3, that is, the proficiency effects, a two-way repeated-measures ANOVA was conducted to compare groups 1 and 2 results across the four time points. Two-way repeated-measures ANOVA is used when there are two factors (G1 and G2) and the same participants receive more than one test (1 to 4 rounds).

4. Results

4.1 *CAF Trajectories*

Complexity. Table 4 and Figure 1 display the results for SC1. For Group 1, SC1 scores decreased slightly from Round 1 to 2, improved from Round 2 to 3, and again decreased from Round 3 to 4. For Group 2, SC1 decreased from Round 1 to 2 and from Round 2 to 3. However, there was an increase from Round 3 and 4.

One-way repeated-measures ANOVA was used to compare the mean scores for each group in Rounds 1, 2, 3, and 4. For Group 1, statistically significant differences were found among the four time points ($df = 3$, $F = 5.48$, $p < .001$, $r = .79$). Post hoc comparison with Bonferroni correction showed significant differences between Rounds 2 and 3 but no significant differences between Rounds 1 and 4 ($p < .50$). For Group 2, mean scores for Rounds 1, 2, 3 and 4 differed statistically significantly among time points (df = 3, $F = 3.76$, $p < .02$, $r = .56$). Post hoc comparison using Bonferroni correction showed significant differences between Rounds 2 and 3 ($p < .001$) but no significant differences between Rounds 1 and 4 ($p < .50$).

Table 4. Means and standard deviations for Syntactic Complexity 1

	Syntactic Complexity 1							
	Round 1		Round 2		Round 3		Round 4	
Group	M	SD	M	SD	M	SD	M	SD
G1 (n=17)	8.03	2.10	7.42	1.20	9.59	2.33	8.36	1.34
G2 (n=14)	8.98	1.97	7.43	1.10	7.40	0.99	8.44	1.46
Mean	8.51	2.04	7.42	1.15	8.50	1.66	8.40	1.40

Figure 1. Trajectories for Syntactic Complexity 1

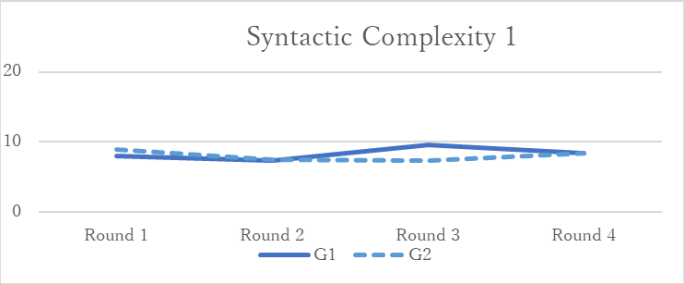


Table 5 and Figure 2 display the results for SC2. For Group 1, the mean SC2 was 1.02 in Round 1, which decreased slightly to 0.99 in Round 2 and improved to 1.39 in Round 3. It fell again to 1.20 in Round 4. For Group 2, the mean score increased linearly (M = 1.03, 1.08, 1.28, and 1.31).

One-way ANOVA was conducted to compare mean scores in Rounds 1, 2, 3, and 4 for each group. For Group 1, it was found that mean scores differed statistically significantly among time points (df = 3, F = 10.05, p < .001, r = .77). Post hoc comparison with Bonferroni correction showed significant differences between Rounds 2 and 3 (p < .001) and between Rounds 1 and 4 (p < .03). ANOVA for Group 2 found that mean scores in Rounds 1, 2, 3, and 4 differed statistically significantly among time points (df = 3, F = 6.901, p < .001, r = .70). A post hoc comparison test with Bonferroni correction showed significant differences between Rounds 1 and 4 (p < .01).

Table 5. Means and standard deviations for syntactic complexity 2

	Syntactic Complexity 2							
	Round 1		Round 2		Round 3		Round 4	
Group	M	SD	M	SD	M	SD	M	SD
G1(n=17)	1.02	0.13	0.99	0.19	1.39	0.31	1.20	0.33
G2(n=14)	1.03	0.20	1.08	0.13	1.28	0.17	1.31	0.23
Mean	1.03	0.16	1.03	0.16	1.33	0.24	1.26	0.28

Figure 2. Trajectories for syntactic complexity 2

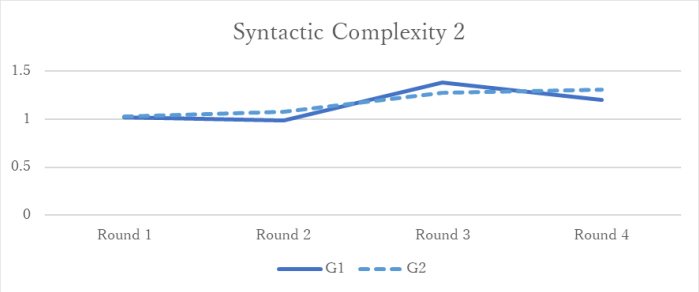


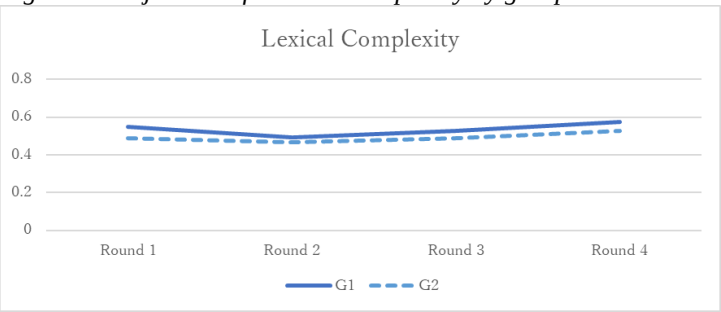
Table 6 and Figure 3 display the LC results. Group 1's mean score decreased slightly from .55 to .49 between Rounds 1 and 2. It went up from .49 to .53 between Rounds 2 to 3. Between Rounds 3 and 4, it rose from .53 to .57. The trajectory for Group 2 was similar to that for Group 1: Round 1 = .49, Round 2 = .47, Round 3 = .49, and Round 4 = .53.

One-way repeated-measures ANOVA was conducted to compare the mean scores in Rounds 1, 2, 3, and 4. For Group 1, the mean scores differed statistically significantly among the time points (df = 3, F = 4.11, p < .01, r = .58). Post hoc comparison using Bonferroni correction showed significant differences between Rounds 1 and 2 (p < .03) but no significant differences between Rounds 1 and 4 (p < .12). For Group 2, the differences between Round 1 and 4 were not statistically significant (p < .32).

Table 6. Means and standard deviations for lexical complexity

	Lexical Complexity							
	Round 1		Round 2		Round 3		Round 4	
Group	M	SD	M	SD	M	SD	M	SD
G1 (n=17)	0.55	0.06	0.49	0.07	0.53	0.07	0.57	0.06
G2(n=14)	0.49	0.13	0.47	0.06	0.49	0.10	0.53	0.05
Mean	0.52	0.10	0.48	0.06	0.51	0.08	0.55	0.06

Figure 3. Trajectories for lexcal complexity by groups

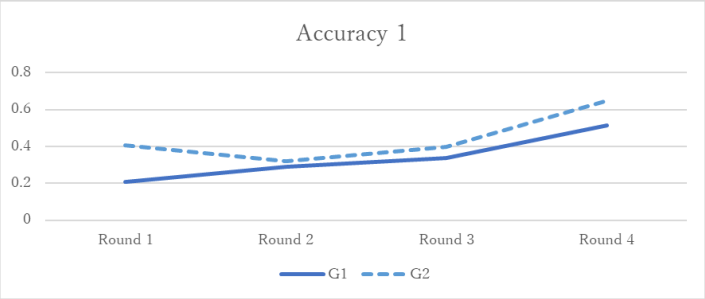


Accuracy. Table 7 and Figure 4 display the results for A1. Group 1's mean scores were .21 in Round 1, .29 in Round 2, .34 in Round 3, and .52 in Round 4. For Group 2, the mean score was .41 in Round 1, .32 in Round 2, .40 in Round 3, and .65 in Round 4.

Table 7. Means and standard deviations for accuracy 1

	Accuracy 1							
	Round 1		Round 2		Round 3		Round 4	
Group	M	SD	M	SD	M	SD	M	SD
G1 (n=17)	0.21	0.19	0.29	0.25	0.34	0.15	0.52	0.20
G2 (n=14)	0.41	0.31	0.32	0.24	0.40	0.11	0.65	0.17
Mean	0.31	0.25	0.31	0.24	0.37	0.13	0.58	0.18

Figure 4. Trajectories for accuracy 1



One-way ANOVA was conducted to compare mean scores for Rounds 1, 2, 3, and 4. For Group 1, mean scores differed statistically significantly among the time points (df = 3, F = 7.18, p < .001, r = .84). A post hoc comparison test using Bonferroni correction showed significant differences between Rounds 3 and 4 (p < .001) and between Rounds 1 and 4 (p < .001). For Group 2, mean scores differed statistically significantly among time points (df = 3, F = .6.50, p < .001, r = .68). Post hoc comparison using Bonferroni correction showed significant differences between Rounds 3 and 4 (p < .001) and between Rounds 1 and Round 4 (p < .01).

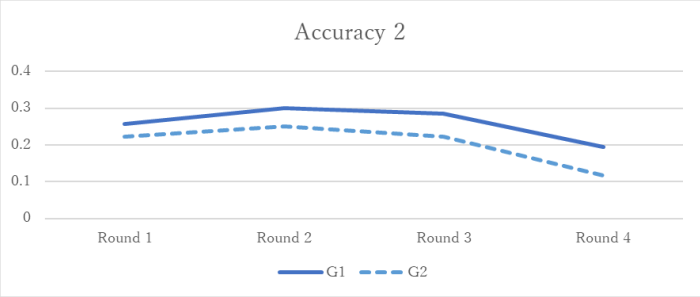
Table 8 and Figure 5 display results for A2. Group 1's mean score was .26 in Round 1, .30 in Round 2, .12 in Round 3, and .20 in Round 4. Group 2's mean score was .23 in Round 1, .28 in Round 2, .26 in Round 3, and .16 in Round 4.

One-way repeated-measures ANOVA was conducted to compare the mean scores in Rounds 1, 2, 3, and 4. For Group 1, mean scores differed statistically significantly among time points (df = 3, F = 4.50, p < .01, r = .78). Post hoc comparison with Bonferroni correction showed significant differences between Round 3 and 4 (p < .001) but no significant differences between Rounds 1 and 4 (p < .07). For Group 2, the mean scores differed statistically significantly among time points (df = 3, F = 6.521, p < .001, r = .83). Post hoc comparison with Bonferroni correction showed significant differences between Rounds 3 and 4 (p < .001) Rounds 1 and 4 (p < .01).

Table 8. Means and standard deviations for accuracy 2

	Accuracy 2							
	Round 1		Round 2		Round 3		Round 4	
Group	M	SD	M	SD	M	SD	M	SD
G1 (n=17)	0.26	0.13	0.30	0.11	0.29	0.12	0.19	0.07
G2 (n=14)	0.22	0.09	0.25	0.12	0.22	0.05	0.12	0.07
Mean	0.24	0.11	0.28	0.11	0.25	0.08	0.16	0.07

Figure 5. Trajectories for accuracy 2

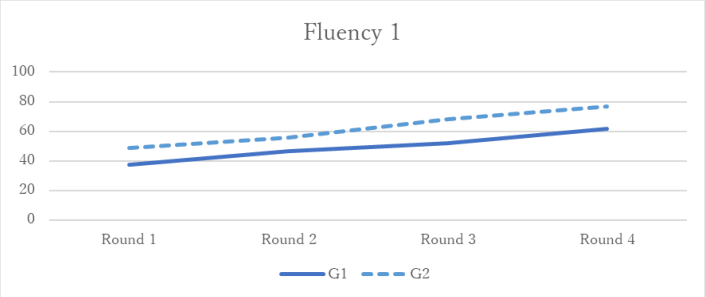


Fluency. Table 9 and Figure 6 present the results for F1. Group 1's mean score in Round 1 was 37.24 and improved by 9.29 words. From Rounds 2 to 3, it rose by 5.65 words. From Rounds 3 to 4, it rose by 9.71 words. Group 2's mean score rose from 48.93 to 55.71 between Rounds 1 and 2, improved to 68.36, and improved to 8.64 words in Round 4.

Table 9. Means and standard deviations for fluency 1

	Fluency 1							
	Round 1		Round 2		Round 3		Round 4	
Group	M	SD	M	SD	M	SD	M	SD
G1 (n=17)	37.24	11.40	46.53	14.95	52.18	14.03	61.88	12.04
G2 (n=14)	48.93	11.00	55.71	9.02	68.36	13.37	77.00	12.71
Mean	43.08	11.20	51.12	11.98	60.27	13.70	69.44	12.37

Figure 6. Trajectories for fluency 1



The results for one-way repeated-measures ANOVA showed that Group 1's mean scores differed statistically significantly among time points (df = 3, F = 10.87, p < .001, r = .89). Post hoc comparison using Bonferroni correction showed significant differences between Rounds 1 and 2 (p < .04), Rounds 3 and 4 (p < .02), and Rounds 1 and 4 (p < .001). Group 2's result for one-way repeated-measures ANOVA showed that mean scores differed statistically significantly among time points (df = 3, F = 18.29, p < .001, r = .93). Post hoc comparison using Bonferroni correction showed significant differences between Rounds 2 and 3 (p < .001), between Rounds 3 and 4 (p < .04), and between Rounds 1 and 4 (p < .001).

Table 10 and Figure 7 display the results for F2. Group 1's mean score was 4.94 in Round 1, 6.29 in Round 2, 7.82 in Round 3, and 9.06 in Round 4. Group 2's mean score was 5.86 in Round 1, 8.36 more in Round 2, 11.86 in Round 3, and 12.00 in Round 4.

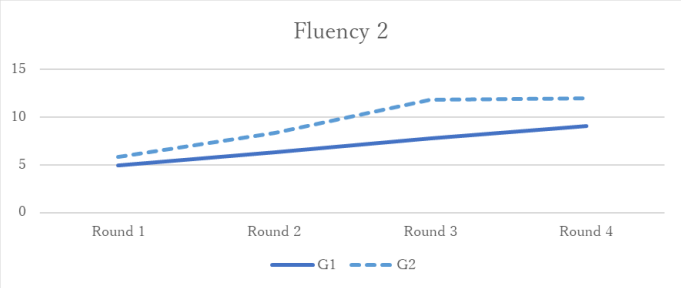
A one-way repeated-measures ANOVA revealed that mean scores differed statistically significantly among the time points (df = 3, F = 8.54, p < .001, r = .86). Post hoc comparison with Bonferroni correction showed significant differences between Rounds 2 and 3 (p < .07) and Rounds 1 and 4 (p < .001). The results of one-way repeated-measures ANOVA for Group 2 showed statistically significant mean score differences among time points (df = 3, F = 25.76, p < .001, r = .95). Post hoc comparison with Bonferroni correction showed

significant differences between Rounds 1 and 2 ($p < .02$), Rounds 2 and 3 ($p < .001$), and Rounds 1 and 4 ($p < .001$).

Table 10. Means and standard deviations for fluency 2

	Fluency 2							
	Round 1		Round 2		Round 3		Round 4	
Group	M	SD	M	SD	M	SD	M	SD
G1 (n=17)	4.94	2.01	6.29	2.52	7.82	2.94	9.06	3.28
G2 (n=14)	5.86	1.88	8.36	2.59	11.86	2.57	12.00	2.04
Mean	5.40	1.94	7.33	2.55	9.84	2.75	10.53	2.66

Figure 7. Trajectories for fluency 2



4.2 *CAF Interaction*

To answer research question 2, within-individual correlation analyses were conducted for groups 1 and 2, the results of which are shown in Tables 11 and 12.

Table 11. Within-individual correlations for CAF in Group 1

	Syntactic Complexity 1	Syntactic Complexity 2	Lexical Complexity	Accuracy 1	Accuracy 2	Fluency 1	Fluency 2
Syntactic Complexity 1	-----	-----	-0.39**	0.05	-0.07	0.20*	0.17*
Syntactic Complexity 2	-----	-----	-0.08	-0.02	-0.11*	0.08	0.50**
Lexical Complexity			-----	-0.47**	0.74**	0.37**	0 .10*
Accuracy 1				-----	-----	-0.29*	0 .10*
Accuracy 2				-----	-----	0.36**	0.02
Fluency 1						-----	-----
Fluency 2						-----	-----

Note: *Significant at p < 0.05 level; **significant at p < 0.01 level

For Group 1, within-individual correlation analyses found a modest negative correlation between LC and SC1 (r = -.39) and between A2 and SC2 (r = -.11). Weak positive relationships were observed between SC1 and F1 and F2 (r = .20; r = .17). SC2 and F2 were strongly correlated (r = .50).

LC had a strong negative correlation with A1 (r = -.47) and a positive one with A2 (r = .74). LC and F1 and F2 showed positive correlations (r = .37; r = .10).

Correlation analysis showed a weak negative relationship between A1 and F1 (r = -.29) but a positive relationship with F2 (r = .10), along with a modest positive relationship between A2 and F1 (r = .36).

Table 12. Within-individual correlations for CAF in Group 2

	Syntactic Complexity 1	Syntactic Complexity 2	Lexical Complexity	Accuracy 1	Accuracy 2	Fluency 1	Fluency 2
Syntactic Complexity 1	-----	-----	-0.24*	0.13*	0.04	0.30**	-0.13*
Syntactic Complexity 2	-----	-----	-0.39**	-0.02	-0.28*	0.04	0.52**
Lexical Complexity			-----	-0.21*	0.36**	0.03	-0.08
Accuracy 1				-----	-----	0.01	-0.07
Accuracy 2				-----	-----	0.22*	-0.01
Fluency 1						-----	-----
Fluency 2						-----	-----

Note: *Significant at p < 0.05 level; **significant at p < 0.01 level

For Group 2, a negative relationship was seen between SC1 and LC (r = -.24) and between SC2 and LC (r = -.39). A weak positive relationship between A1 and SC1 (r = .13) and a negative one was seen between SC2 and A2 (r = -.28). A modest positive correlation was seen between SC1 and F1 (r = .30) but a weak negative one between SC1 and Fluency 2 (r = -.13). SC2 and F2 were strongly correlated (r = .52).

LC had a weak negative correlation with A1 (r = -.21) and a mild positive relationship with A2 (r = .36). No meaningful result was seen between LC and F1 or F2.

The analysis showed a mild correlation between A2 and F1 (r = .22). There was no correlation between A1 and F1 or F2.

4.3 *Proficiency Effects*

To examine research question 3, a two-way repeated-measures ANOVA was conducted. Table 13 shows significant differences between Groups 1 and 2 in SC1 (df = 3, F = 5.61, p < .001, r = .81). On other measures, no statistically significant differences were seen between the two groups.

Table 13. Differences between Group 1 and 2 across the four time points

	F	*p*	*r*
Syntactic Complexity 1	5.61	0.00	0. 81
Syntactic Complexity 2	1.52	0.21	0.58
Lexical Complexity	0.26	0.86	0.28
Accuracy 1	1.05	0.38	0.28
Accuracy 2	0.33	0.80	0.51
Fluency 1	78.97	0.54	0.32
Fluency 2	13.88	0.61	0.98

Note: *df* = 3

4.4 *Self-reported Evaluation (Questionnaire)*

A seven-item questionnaire was used to explore other factors that might affect the results of the above CAF data. Tables 14 and 15 display the results of students' self-reported evaluations.

Table 14. Self-reported evaluation for Group 1

	Round 1		Round 2		Round 3		Round 4	
	M	SD	M	SD	M	SD	M	SD
1. This task was difficult.	3.26	1.10	2.95	1.03	2.94	1.00	3.37	1.01
2. I was able to express my opinion.	2.74	0.87	3.11	0.81	3.39	0.78	3.47	0.77
3. I paid attention to correct pronunciation while speaking.	2.89	0.94	2.84	0.69	3.17	0.71	3.63	0.90
4. I paid attention to correct grammar usage while speaking.	2.47	0.77	2.58	0.61	2.83	0.79	3.37	0.68
5. I found myself thinking in Japanese when delivering a speech.	2.89	1.56	2.47	1.47	2.50	1.42	2.37	1.26
6. I felt confident when I was speaking English during the task.	2.00	1.11	2.32	0.67	2.56	0.62	2.89	0.74
7. I didn't worry about making mistakes during the task.	2.95	0.97	3.21	1.08	3.33	0.97	3.42	1.02

Table 15. Self-reported evaluation for Group 2

	Round 1		Round 2		Round 3		Round 4	
	M	SD	M	SD	M	SD	M	SD
1. This task was difficult.	3.05	0.91	3.06	0.80	2.89	0.96	3.11	1.20
2. I was able to express my opinion.	2.79	0.98	3.44	0.78	3.44	0.92	3.58	0.96
3. I paid attention to correct pronunciation while speaking.	2.79	1.18	2.89	1.02	2.83	0.79	3.05	0.78
4. I paid attention to correct grammar usage while speaking.	2.58	1.07	2.72	0.96	3.06	0.94	3.26	0.73
5. I found myself thinking in Japanese when delivering a speech.	2.05	1.31	2.44	1.10	1.89	1.13	2.21	1.23
6. I felt confident when I was speaking English during the task.	2.53	1.17	2.89	0.96	3.11	0.76	3.21	1.03
7. I didn't worry about making mistakes during the task.	3.42	1.22	3.50	1.29	3.89	1.13	3.95	1.03

Group 1 students felt that the task became easier and that they were better able to convey their messages as time passed, although one-way ANOVA found that these differences were not significant (df = 3, F = .89, p < .45, r = .48; df = 3, F = 2.51, p < .07, r = .68). Level of attention to pronunciation dropped slightly between Rounds 1 (M = 2.90) and 2 (M = 2.84) but rose gradually from Rounds 2 to 4 (M = 3.63). One-way ANOVA showed that differences were statistically significant (df = 3, F = 2.98, p < .04, r = .71), but post hoc comparison with the Bonferroni correction showed no significant differences between any pair of time points. Attention to syntax improved significantly, with positive linear growth from Rounds 1 to 4 (df = 3, F = 2.61, p < .001). Post hoc comparison with the Bonferroni correction showed significant differences between Rounds 1 and 4 (p < .01) and between Rounds 2 and 4 (p < .01). The ability to retrieve correct English words was nonlinear and showed no statistically significant differences (df = 3, F = 1.22, p < .62, r = .54). Students' confidence in speaking English improved significantly on a positive linear trajectory as mean scores improved over time (2.00–2.32–2.56–2.90). One-way ANOVA found that the differences were statistically significant (df = 3, F = 2.05, p < .04, r = .64), but post hoc comparison using the Bonferroni correction showed no significant differences between any of the time points. The students' scores for anxiety about making mistakes improved as well, from 2.95 in Round 1 to 3.21 in Round 2, 3.33 in Round 3, and 3.42 in Round 4. This time, however, the differences were not statistically significant (df = 3, F = .43, p < .73, r = .36).

The results for the two groups were similar. Group 2 found the task least difficult at Round 3, but the differences were not statistically different (p < .87, r = .27). The attention to context improved over time in a statistically significant pattern (df = 3, F = 3.04, p < .04, r = .71), tracing positive linear growth, but post hoc comparison with Bonferroni correction presented no significant differences for any time point. The attention to pronunciation of both Groups 1 and 2 was nonlinear, moving up and down without significant differences (df = 3, F = .31, p < .82, r = .31). As time passed, students paid more attention to grammar, as is seen in the growth of mean scores (2.58–2.72–3.06–3.26); the difference was not statistically significant (df = 3, F = 2.68, p < .05, r = .69). Both groups' developmental pattern for retrieving English words was nonlinear; ANOVA indicated no significant difference (df = 3, F = .80, p < .50, r = .46). Confidence in speaking English and level of anxiety about making mistakes showed positive linear growth over time. One-way ANOVA indicated that the difference was not significant (p < .27, r = .56; p < .41, r = .50).

5. Discussion

5.1 *CAF trajectories*

The first research question concerned developmental patterns in CAF. The examination of SC1 found Group 1 with a nonlinear trajectory characterized by fluctuations, and Group 2's trajectory had a mild U-shape curve. However, according to the results of the post hoc test, there were no differences for SC1 in either group, except between rounds 2 and 3, when G1's score improved but G2's score decreased. Some ups and downs were seen for Group 1 and some changes for Group 2; however, except between Rounds 2 and 3, these rates were too small for a change in competency to be established for either group for SC1. The differences between Rounds 2 and 3 could have been because of the prompt type because, for the first, second, and fourth prompts, the speakers were asked to compare two options, whereas, for the third prompt, the speakers had to answer a simple yes/no question without any options. Therefore, the speakers, possibly making them perform differently or produce different utterances, might have perceived the third prompt differently.

The measurement for the rate of clauses per AS-unit (SC2) indicated some improvements in both groups. Although they showed slightly different trajectories, over time, the students in both groups were able to make more complex sentences.

Most differences were not statistically significant, but both trajectories (SC1 and SC2) showed mild fluctuations for Group 1. To identify factors that could have affected the results, students' self-evaluations were examined. In the Round 1 open-ended comment box, one student wrote (translated by the Author), "There are many things I wanted to say, but I could not explain them in detail." Then, in Round 2, she wrote, "Compared to Round 1, I think I did better because I just tried to push out all my thoughts as much as possible." In Round 3, she stated, "I wasn't able to express what I wanted to say. The sentences are always short. I had a hard time." Finally, in Round 4, she explained, "This is very difficult, and I'm not good at it. To overcome anxiety, I would like to continue and do whatever I can." This student's perceived competence was not consistent over time but showed some fluctuations. As Norris and Ortega (2009) showed, learners' willingness to communicate and their performance may be correlated. Overall, a mild positive growth was seen in syntactic complexity for both groups.

The change trajectories in lexical complexity for both groups showed a U-shaped curve. The overall development was mild, and slight changes in growth were observed. No significant changes over time were seen in the self-evaluation of the retrieval of lexical items. The mean scores remained about the same for both groups, meaning that the level of effort for retrieving varied lexical items remained the same over time. Thus, lexical complexity may not be related to the retrieval of lexical items but in some way to learners' lexical repertoires.

Students' accuracy showed a U-shaped trajectory for both groups, with a decline followed by a steep increase. The scores for anxiety regarding making mistakes improved over time (in Round 1, students were more afraid of making errors than they were later). In the early stages, therefore, students might have avoided making mistakes due to anxiety. Over time, as this anxiety decreased, they produced more sentences with content, improving their content scores, leading to more errors than before. Still later, however, the number of errors decreased. As their error anxiety scores were lower in the later stages, it is possible that an increased scope could be seen for attention to grammar. The data from this study do not support Seo and Eo (2011), who found that as proficiency improved, accuracy declined.

In addition, the developmental path the learners followed was not consistent with that in Larsen-Freeman's (2006) longitudinal study, where more irregularity was found in accuracy than fluency or complexity. Larsen-Freeman argues that the "development of accuracy is not discrete and stage-like but more like the waxing and waning of patterns" (Larsen-Freeman, 2006, p. 590).

Both F1 and F2 showed a positive linear change for the two groups. The questionnaire results indicate that students' confidence in speaking significantly improved on a positive linear trajectory, with mean scores improving over time (2.00–2.32–2.56–2.90). These data show a perfect positive correlation between the developmental path of fluency and students' confidence in speaking (i.e., higher confidence scores mean higher fluency scores). As McIntyre and Gardner (1994) indicate, foreign language anxiety has a strong connection with L2 learning. According to Totb (2014), fluency is the most conspicuous speech characteristic that distinguishes L2 learners with a high level of anxiety from those with a lower one. In addition, because fluency is defined by how fast the learner speaks without showing dysfluency markers (Wolfe-Quintero et al., 1998), many studies (e.g., Freed, 2000; Koizumi, 2005; Kormos & Dénes, 2004; Skehan & Foster, 1999) consider disfluency markers to indicate lack of fluency. However, the number of dysfluency markers observed did not show the same developmental pattern as learner fluency in this study, which appears to contradict the trade-off claimed by Robinson (2001, 2003, 2005) and Yuan and Ellis (2003). Given the U-curved development patterns (Group 1: 3.63–6.39–5.89–5.62; Group 2: 7.56–8.94–13.79–11.35) correlated with accuracy patterns in this study, dysfluency markers should be considered possible markers of accuracy.

5.2 *CAF Interaction*

This study found lexical complexity was negatively correlated with syntactic complexity and accuracy for both groups but positively correlated with fluency, contradicting Skehan (1996), Michel et al. (2007), and Vercellotti's (2017), among others. According to Michel et al. (2007), complex tasks generate more accuracy and lexical complexity but not grammatical complexity. This study supports their view of the relationship between lexical richness and syntactic complexity. More attention to lexical items probably led to longer and faster utterances but also less complex and accurate production. In Vercellotti's (2017) longitudinal study, lexical variety was positively correlated with accuracy, grammatical complexity, and fluency. She also claimed that lexical variety could measure learners' general proficiency. However, here, lexical complexity was negatively correlated with other CAF components, except for fluency. Moreover, according to McWhinney (2001), as lexical items are activated and retrieved before sentence production rather than during sentence formulation, the retrieval of varied lexical items should not affect syntactic complexity. In the results of learners' vocabulary retrieval from the self-evaluation questionnaire, the perception of the difficulty of word retrieval did not change over time. Thus, lexical complexity appears to be more closely correlated with the learners' lexical repertoire than with their retrieval. In any case, lexical complexity was found here to be negatively correlated with syntactic complexity and accuracy for both groups, meaning that the more that a learner's lexical variety increased, the more likely it was that the learner would fail at or abandon the production of complex and accurate sentences but gain fluency nevertheless.

Robinson (1995, 2001, 2003, 2005) found that complexity and accuracy are correlated and should increase or decrease simultaneously. However, other researchers (e.g., Bygate, 2001; Skehan and Foster, 1997) found that complexity and accuracy compete with

one another. Benevento and Storch (2011) observed improvements in language complexity of writing in secondary school L2 learners but no significant improvements in accuracy over some time. This study partially supports Robinson's hypothesis. In particular, A1 was positively correlated with syntactic complexity, and A2 had a negative correlation with syntactic complexity. In other words, as learners became able to produce more accurate outputs, their utterances became increasingly complex; that is, they produced more complex utterances with no loss of accuracy.

According to Robinson (1995, 2001, 2003, 2005), accuracy and complexity progress at the expense of fluency. Skehan (1996), however, suggests that these variables compete for attentional resources. The results of this study, however, differ from these findings. Although F2 in Group 2 had a weak negative correlation with SC1, a strong tendency for a positive correlation was seen between fluency and complexity, accompanying a highly significant relation between F2 and SC2 in both groups. F1 and SC1 were also positively correlated for both groups, meaning that learners were likely to use more complex structures as they produced more words.

The results for the relationship between fluency and accuracy were mixed. For Group 1, there was a weak negative correlation between F1 and A1 and a mild positive correlation between F1 and A2, indicating that the more that learners spoke, the more likely they were to make errors. Conversely, however, a weak positive correlation was seen between F2 and A1 for the same group. Moreover, in Group 2, a positive correlation was observed between F1 and A2, meaning that the more words the learners produced, the greater their inaccuracy. Although it might be thought that learners cannot be fluent and accurate at the same time, these results indicate that there is no strict trade-off, as claimed by Robinson (2001, 2003, 2005) and Yuan and Ellis (2003). At the same time, we cannot support Vercellotti's (2017) conclusions, as our results were insufficiently clear, and the relationship between fluency and accuracy did not reach significance. To better understand the relationship between fluency and accuracy, results with more dynamic descriptions are required (Larsen-Freeman, 2006). For example, more detailed aspects of performance, considered via an in-depth analysis of quantitative data focusing on individual differences, may help identify the relationship more accurately. Kormos (1999) reported a mixture of qualitative and quantitative observations using interviews and questionnaires to determine whether differences in individual speaking tendencies such as "monitor under-user" or "monitor over-user"[vi] were reflected in differences in oral production. She concluded that monitor over-users spoke much less fluently and rephrased themselves more than monitor under-users. Moreover, monitor over-users used disfluency markers such as self-correction less frequently. According to the self-evaluations in this study, decreased anxiety regarding the commission of errors resulted in more utterances and errors over time. Learners' scores for attention to grammar usage also improved over time; however, this resulted in additional dysfluency markers with self-corrections and more utterances. These results suggest that researchers should not neglect the impact of individual differences.

5.3 *Proficiency Effects*

The third research question investigated the proficiency effect, namely, whether learners' proficiency levels influence CAF development. Do learners with lower proficiency levels have steeper rates of improvement because of their larger room for growth, or do they experience cognitive overload, which slows their growth? It was found here that the CAF developmental pattern for lower proficiency students was identical to that of higher proficiency students, except in the case of SC1, which indicated a salient difference in Round

3 for both groups. The complexity score for Group 2 (Intermediate) decreased in Round 3, but it remained about the same for Group 1.

This finding may be caused by the degree of cognitive task complexity owing to the results of the self-evaluation questionnaire. The mean task difficulty score in Round 3 decreased for the intermediate-level students (i.e., the students did not find the task as difficult). Robinson's cognition hypothesis proposed that a cognitively demanding task would result in greater complexity in L2 production. According to Ishikawa (2007), the participants involved in a more complex task (−here and −now) obtained higher performance in syntactic scores than the participants involved in a less complex task. Therefore, decreased syntactic complexity in learners' L2 production could have occurred in response to a relatively less demanding task.

The results of the current study suggest that L2 learners at low and intermediate proficiency levels follow the same pattern of CAF development. At the same time, there may be a relationship between CAF development and task complexity. To that end, it is important to take proficiency effects and task types into consideration when observing CAF development, because proficiency may interact with the cognitive demands imposed by task complexity (Robinson, 2005; Sasayama, 2016). Sasayama (2016) suggested that learner proficiency mediates cognitive complexity; learners at different proficiency levels devote the same attentional resources to performing the same tasks, but they both perform them differently and perceive the tasks differently due to the difference in cognitive load.

6. Conclusion

This study examined developmental patterns in oral language performance as measured by CAF and the relationship between these three variables. Proficiency effects of longitudinal improvement in CAF were also examined. The participants were 31 Japanese-L1 university students. An impromptu speech task was administered four times, and the students' speech was recorded, transcribed, coded, and analyzed to measure CAF.

Complexity saw a mild growth over time, with some fluctuations. Accuracy had U-shaped trajectories, and fluency grew linearly over time. Moreover, some results concerning the correlations among CAF were congruent with those of previous empirical research, while others were not. Widely known trade-off effects between lexical and syntactic complexity and between lexical complexity and accuracy were evident, as predicted by Skehan's limited capacity hypothesis (1988). However, accuracy and syntactic complexity were correlated, supporting Robinson (2001, 2003, 2005). However, these results partially refute Robinson, as fluency and syntactic complexity displayed positive correlations. The relationship between fluency and accuracy was unclear. It was suggested that other factors, such as individual differences may affect learners' oral production. For the proficiency effect, it was found that both student groups had similar change trajectories, except for SC1.

This study had limitations that could be compensated for by further research. First, the number of participants was relatively small, requiring a large-scale study to confirm the findings. Second, the research was limited to learners at the lower and intermediate levels, so other levels should be investigated. Third, the overall findings may have been affected by other factors, such as the curriculum, pedagogical approach, and type of task. Fourth, although the observations spanned 3 months, even longer observation periods could yield different change trajectories that merit further research. Fifth, this study adopted seven measures; however, some other measures, such as mean length of pause [MLP] for fluency,

should be used to enhance the findings of this study. Finally, the impact of individual differences remains little known. Although quantitative data and traditional statistics may provide figures that are useful for identifying overall tendencies, it is necessary to conduct more in-depth, detailed observations using interview methods to understand L2 learners' CAF development. Individual learners' experiences may differ from those that would be expected from typical development patterns.

[i] The participants were from the Department of Airlines and studying different aspects of the airline industry. They were split into two proficiency groups at the beginning of the semester on the basis of their TOEIC L&R scores. While the two groups attended different classes, they learned the same content from the same teacher using the same textbook. The school curriculum design limited the number of students in each class to no more than 30; therefore, because there were more than 50 students in the junior year, the school split them into two classes based on their lower and intermediate proficiency levels. Although their levels differed, the students' English language learning backgrounds were similar; they had studied English as a subject in junior high and high school for 6 years mostly through the grammar-translation method and audiolingualism. In the college, the students followed the same curriculum for the two-year course; however, there was no English class in the first year. In the second year, a speaking class was offered, the only available English class.

The lessons were conducted by a Japanese bilingual teacher using both English and Japanese; however, the students were encouraged to speak English as much as possible in the class. The teacher conducted various tasks in each class (see Appendix 1 for detail). As the aim of the course was improving the students' overall speaking skills, the teacher planned the lessons, such that students could improve all CAF components; that is to say, the teacher did not explicitly focus on one CAF component.

Pair and group work activities were used throughout the course to alleviate the students' English speaking anxiety, and as a move away from the traditional evaluation system, there were no midterm or paper exams so that the learners could overcome their fear of speaking. Students were given a speech topic homework assignment every other week throughout the term and were given two weeks to write a script and practice delivering the speech at home; however, this was the only time the use of English outside the class was encouraged.

[ii] The Test of English for International Communication (TOEIC) Listening/Reading (L&R) is a 180-minute English language listening and reading test run by the Educational Testing Service (ETS). The TOEIC has been widely used for recruitment, training, and student placement in Japan. Although the TOEIC L&R is designed to measure listening and reading abilities, previous studies (Liu & Costanzo, 2013; Koizumi, 2015; Kanzaki, 2020) found a significant correlation between TOEIC L&R test scores and TOEIC Speaking test scores (a computer-based test to measure speaking skills).

[iii] The FLCAS was referred to in this study because it has been considered the most reliable and valid method to have been used by many researchers (e.g., Horwitz et al., 1986; Tallon, 2009). Original version consists of 33 items, but the questionnaire for this study consists of 7 items because the questionnaire was conducted during regular class time, along with the speaking test, and it was assumed that investing too much time in the questionnaire was not a good idea. In addition, although individual differences, of course, can affect L2 performance, this study mainly considers the CAF development using quantitative evaluation.

[iv] Please note that a question related to the self-report evaluation (questionnaire) was not included as a research question for the following reasons. 1. Although individual differences, such as the anxiety associated with making errors, can affect L2 performance, this study was focused on quantitatively evaluating CAF development. 2. Given the multiple variables (time and proficiency level) that could have affected the dependent variable, adding another factor would further complicate the study and was thus eschewed; that is to say, the method and analysis may have become too complicated. 3. It could be argued that oral performance development can only be accurately measured by considering

individual differences; however, observing these relationships was beyond the scope of this study. Therefore, it was considered more suitable to use a questionnaire to gain supplementary information that would complement the data interpretation. This study did not specifically examine the questionnaire validity for the same reason.

[v] The clausal definitions were based on Foster et al. (2000). An AS-unit (analysis of speech unit) refers to an utterance consisting of an independent clause or a sub-clausal unit plus any subordinate finite or nonfinite clauses (Foster et al., 2000). The analysis allows for the isolation of one or more phrases without a verb that could be elaborated as a full clause with communicative value.

The following are several examples:

1. Where did you put the book? (1 clause, 1 AS-unit)
2. On the table. (0 clause, 1 AS-unit)
3. Because it is expensive. (1 clause, 1 AS-unit)
4. You should stop fooling around. (2 clauses, 1 AS-unit)
5. Yesterday, book on the table. Today, book in the bag. (0 clause, 2 AS-unit)
6. I like this book because it is interesting. (2 clauses, 1 AS-unit [this can be counted as two AS-units depending on intonation and pausing.])

Various recent spoken data studies have used AS-unit analyses because they are applicable to the complex realities in L2 learners' oral transcripts. This study adopted AS-unit analyses for this reason.

[vi] These are notions from Krashen (1978): Monitor under-users tend to be concerned with speed and fluency but not errors. On the contrary, monitor over-users tend to be concerned with form, which, when coupled with anxiety, impedes fluency.

References

Alavi, S. & Sadeghi, K. (2017). Development of Fluency, Accuracy, and Complexity in Productive Skills of EFL learners across Gender and Proficiency: A Chaos Complexity Approach. *Journal of Teaching Language Skills*, *35*(4), 1-35.

Bamanger, E., & Gashan, A. (2014). The effect of planning time on the fluency, accuracy, and complexity of EFL learners' oral production. *Journal of Educational Sciences, 27* (1), 1-15.

Bei, G. X. (2013). Effects of Immediate Repetition in L2 Speaking Tasks: A Focused Study. *English Language Teaching*, *6*(1), 11-19.

Benevento, C., & Storch, N. (2011). Investigating writing development in secondary school learners of French. *Assessing Writing, 16*(2), 97-110.

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing*(pp.23-48). London: Longman.

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness, in the spontaneous speech of bilinguals. *Applied Linguistics, 24*, 197-222.

de Bot, K. (2007). A Dynamic Systems Theory approach to second language acquisition. *Bilingualism: Language and Cognition*, *10*(1), 7-21.

de Bot, K. (2008). Introduction: second language development as a Dynamic Process. *The Modern Language Journal*, *92*, 166-178.

de Bot, K., & Larsen-Freeman, D. (2011). Researching second language development from a dynamic systems theory perspective. In M. Verspoor, K. de Bot, & W. Lowie (Eds.), *A dynamic approach to second language development: methods and techniques* (pp. 5-23). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Ellis, R. & Barkhuizen, G. (2005). *Analyzing learner language*. Oxford: Oxford University Press.

Ellis, R., & Yuan, F. (2005). The effects of careful within-task planning on oral and written task performance. In R. Ellis (Ed.), *Planning and task performance in a second language*(pp.167-192). Amsterdam: John Benjamins.

Ferrari, S. (2012). A longitudinal study of complexity, accuracy and fluency variation in second language development. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. Philadelphia: John Benjamins.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*, 299-323.

Foster, P. & Skehan, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1*(3). 185-211.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21, 354-375.

Freed, B. F. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder? In Heidi Riggenbach (Ed.), *Perspectives on Fluency* (pp.243–265). Ann Arbor: University of Michigan Press.

Griffiths, R. (1991). Pausological research in an L2 context: A rationale, and review of selected studies. *Applied Linguistics*, 12, 345-364.

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern language journal*, *70*(2), 125-132.

Ishikawa, T. (2007). The effect of manipulating task complexity along the [+/−Here-and-Now] dimension on L2 written narrative discourse. *Investigating tasks in formal language learning* (pp.136-156). Clevedon: Multilingual Matters.

Ishikawa, T. (2015). The Influences of Learners' Basic Attributes and Learning Histories on L2 Speech Fluency: A Case Study of Japanese and Chinese Learners' of English. *Social and Behavioral Sciences*, 192, 516-525.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: a developmental perspective. Lund University, Dept. of *Linguistics and Phonetics Working Papers*, *53*, 61–79.

Kanzaki, M. (2020). TOEIC Listening and Reading test and overall English ability. In P. Clements, A. Krause, & R. Gentry (Eds.), *Teacher efficacy, learner agency* (pp.559-567). Tokyo: JALT.

https://doi.org/10.37546/JALTPCP2019-63

Koizumi, R. (2005). Speaking performance measures of fluency, accuracy, syntactic complexity, and lexical complexity. *JABAET (Japan-Britain Association for English Teaching) Journal*, *9*, 5-33.

Koizumi, R. (2015). Factor structure and four-skill profiles of the TOEIC test among Japanese university learners of English. *ARELE: Annual Review of English Language Education in Japan*, 26, 109-124.

Koizumi, R., & In'nami, Y. (2014). Modeling Complexity, Accuracy, and Fluency of Japanese Learners of English: *A Structural Equation Modeling Approach. JALT Journal, 36*(1), 25-46.

Koizumi, R., & Yamanouchi, I. (2003). Nihonjin chuugakusei no speaking no hattatsu [Development in speaking ability among Japanese junior high school students: Using self-introduction task]. *Bulletin of the Kanto-Koshin-Etsu English Language Education Society*, 17, 33-44.

Kormos, J. (1999). The effect of speaker variables on the self-correction behavior of L2 learners. *System*, *27*, 207-21.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*, 145-164.

Kuiken, V., & Vedder, I. (2012). Syntactic complexity, lexical variation and accuracy as a function of task complexity and proficiency level in L2 writing and speaking. In A. Housen, F. Kuiken & I. Vedder (Ed.), *Dimensions of L2 performance and proficiency: complexity, accuracy and fluency in SLA* (pp. 143-169). Amsterdam: John Benjamins.

Krashen, S. (1978). Individual variation in the use of the monitor. In W. Ritchie (Ed.), *Second Language acquisition research: Issues and implications* (pp. 175-83). New York: Academic Press.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in thee oral and written production of five Chinese learners of English. *Applied Linguistics*, *27*(4), 590-619.

Larsen-Freeman, D. (2009). Adjusting expectations: the study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 579-89.

Larsen-Freeman, D., & Long, M. H. (1991). *An Introduction to Second Language Acquisition Research.* Harlow: Longman Group.

Liu, J., & Costanzo, K. (2013). The relationship among TOEIC listening, reading, speaking, and writing ski11s. In D, E. Powers (Ed.), *The research foundation for the TOEIC Test: A compendium of studies: Vol. 2* (pp. 2.1-2.25). Princeton, NJ: Educational Testing Service. https://www.ets.org/Media/Research/pdf/TC2-02.pdf

Malvern, D. D., Richards, B.J., Chipere, N., & Duran, P. (2004). Lexical diversity and language development: Quantification and assessment. Hampshire, England:

Palgrave Macmillan.

McIntyre, P.D., & Gardner, R. C. (1994). The subtle effects of language anxiety on cognitive processing in the second language. *Language Learning*, *44*, 283-305.

McWhinney, B. (2001). The competition model: the input, the context, and the brain. In P. Robinson (Ed.), *Cognition and Second Language Instruction*. Cambridge: Cambridge University Press.

Michel, M. C., F. Kuiken, F. & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics, 45*(13), 241-59.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555-578.

Polat, B., & Kim, Y. J. (2014). Dynamics of complexity and accuracy: A longitudinal case study of advanced untutored development. *Applied Linguistics, 35*(2), 184-207.

Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning, 45*, 99-140.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring Interactions in a componential framework. *Applied Linguistics, 22*(1), 27-57.

Robinson, P. (2003). The cognition hypothesis, task design, and adult task-based language learning. *Second Language Studies 21*(2), 45-105.

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics, 43*, 1-32.

Robinson, P. (2011). Task-based language learning: A review of issues. *Language learning, 61*(1), 1-36.

Sasayama, S. (2016). Is a 'Complex' Task Really Complex? Validating the Assumption of Cognitive Task Complexity. *The Modern Language Journal, 100* (1), 231-254.

Sasayama, S. & Izumi, S. (2012). Effects of task complexity and pre-task planning on Japanese EFL learners' oral production. IN A. Shehadeh & C. A. Coombe (Ed.), *Task-based language teaching in foreign language contexts: Research and implementation* (pp. 23-42). Amsterdam: John Benjamins.

Seo, S. J., & Eo, J. H. (2011). Study on the accuracy variation of connective endings by Korean proficiency level. *Journal of Korean Language Education, 22*(1), 123-143.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics, 17*(1), 38-62.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press. Skehan, P. (2015). Limited attention capacity and cognition: Two hypothesis regarding second language performance on tasks. In M. Bygate (Ed.), *Domains and Directions in the Development of TBLT* (pp.123-155). Amsterdam: John Benjamins.

Skehan, P. & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1*(3), 185-211.

Skehan, P. & Foster, P. (1999). The Influence of Task Structure and Processing Conditions on Narrative Retellings. *Language Learning, 49*(1). 93-120.

Spoelman, M. and M. Verspoor. (2010). Dynamic patterns in development of accuracy and complexity: a longitudinal case study in the acquisition of Finnish. *Applied Linguistics, 31*(4), 532-53.

Takiguchi, H. (2004). Nihonjin EFL chuugakusei no speaking nouryoku no hattatu kenkyu [A study of the developments of speaking ability among Japanese junior high school students from the viewpoints of fluency, complexity, and accuracy]. *KATE (Kanto-koshinetsu Association of Teachers of English) Bulletin, 18*, 1-13.

Tallon, M. (2009). Foreign Language Anxiety and Heritage Students of Spanish: A Quantitative study. *Foreign Language Annals, 42*(1), 112-137.

Totb, Z. (2014). A Native Speaker's Perceptions of High VS. Low Anxious EFL Students' Speaking Performance. In J. Horvath & P. Medgyes (Ed.), *Studies in Hour of Marianne Nikolov* (pp.259-273). Pecs: Lingua Franca Csoport.

Vercellotti, M. L. (2017). The Development of Complexity, Accuracy, and Fluency in Second Language Performance: A Longitudinal Study. *Applied Linguistics, 38*(1), 90-111.

Verspoor, M., Lowie, W., & Van Dijk, M. (2008). Variability in second language development from a dynamic systems perspective. *Modern Language Journal, 92*, 214-231.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy & complexity*. Honolulu, hi: University of Hawai'i Press.

Yang, W. & Sun, Y. (2015). Dynamic Development of Complexity, Accuracy and Fluency in Multilingual Learners' L1, L2 and L3 Writing. *Theory and Practice in Language Studies, 5*(2), 298-308.

Yuan, F & Ellis, R. (2003). The effects of pre-task planning and online planning on fluency, complexity, and accuracy in L2 monologue oral production. *Applied Linguistics, 24*(1), 1-27.

Appendix 1: Class syllabus

Class	Date	Class schedule	Homeword	Speaking Test
Class 1	11-Apr	Introduction Ice-breaking activities/Introducing yourself		
Class 2	18-Apr	TOEIC Speaking Sample Test/Pronunciation Practice Section 1: Reading Alound/criteria/recording&assessing	Give Homework 1	Round 1
Class 3	25-Apr	Homework check (speech) Section 1 review/ Best Announcer Award Pausing, intonation, assiilation, linking, fragments		
Class 4	9-May	Section 2: Picture description/criteria/recording&assessing	Give Homework 2	
Class 5	16-May	Homework check (speech) Section 2 reveiw Retelling the picture, Concentration, Best Picture Game		Round 2
Class 6	23-May	Section 3: Respond to questions/criteria/recording&assessing	Give Homework 3	
Class 7	30-May	Homework check (speech) Section 3 Review/Quick Response QA/Interview game(pair&group)		
Class 8	6-Jun	Section 4:Make your own chart & create own qestions/criteria/recording&assessing	Give Homework 4	
Class 9	13-Jun	Homework check (speech) Section 4 Review/Make your own chart & create own qestions		Round 3
Class 10	20-Jun	Section 5: Propose a solution/criteria/recording&assessing	Give Homework 5	
Class 11	27-Jun	Homework check (Speech) Sction 5 review/ Proposal chart,Giving advice game (cell phones), Retelling game		
Class 12	4-Jul	Section 6: Making a speech/criteria/recording&assessing	Give Homework 6	
Class 13	11-Jul	Homework check (speech) Sample TOEIC Speaking Test (Speaking Computer Lab) Discuss your performance as a group		
Class 14	18-Jul	Group Interviews		Round 4